Smartphone Market Data Analysis using Data Warehouse

Brenden Carvalho, Siddhaling Urolagin
Department of Computer Science
BITS Pilani, Dubai Campus, Dubai International Academic City
PO Box 345055, Dubai

1. Introduction

Smartphones are the primary configuration of mobile communication devices and currently represent the fastest growing segment in the telecom industry. Apple entered the market in 2007 by developing the first model of iPhone. The iPhone revolutionized the smartphone industry by offering a full touchscreen display and using a virtual keyboard. The first android phone on the other hand wasn't released until late 2008. In 2013, for the first-time sales of internet-connected smartphones exceeded that of more basic handsets. Since then the competition in the market has been fierce among incumbents and between incumbents and new entrants into the market, each trying to win over the other for market share by offering their products with lower prices and therefore lower margins while continuing to improve the standard and quality of their devices. The smartphone industry has been growing steadily in both market size and manufacturers. Global smartphone shipments are projected to reach a total of 1.48 billion in 2023. Further, by the end of 2020, almost 44.9 percent of the world's population will own or use a smartphone.

With so many smartphone users, every smartphone brand, large or small will have large scale chronological data available from their day-to-day operations. This might include sales, logistics, development, device usage behavior, component specification and other operational data. The objective of this work is to use this data to develop a data warehouse model for the smartphone market. This could then be used to develop a prediction or analytic model that could support the smartphone brand's decision-making process on matters such as budget allocation, product development, feature development as well as other aspects such as sales forecasting, market prediction among others.

The paper is structured as follows: In section 2, we present some of the related work performed in this field. In section 3, we discuss about the architecture implemented for this paper. In section 4, we outline the ETL process. In section 5, we discuss about MDX queries that are used and the results obtained. And finally, in section 6, we summarize our findings and conclude.

2. Literature Survey

Data warehouse has been applied in the smartphone industry before, in [1], the authors utilize the telecom operator's call detail record to develop a data warehouse. Analytical processing is carried out using the data warehouse. A lattice of cuboids is constructed to carry out the OLAP processing from all possible business perspective. This helps the authors

to understand the market share of different smartphone vendors among other aspects. Data warehouse techniques are commonly used for business intelligence. In [2], the authors propose a bitemporal spatio-temporal cloud-based data warehouse for business intelligence for a telecommunications company. The starnest logical schema was used for the data warehouse which is a combination of star and snowflake schemas. The starnest schema was further expanded to support spatial and temporal data using subtuples. Data warehouse can be used for other purposes besides business intelligence as well. In [3], the authors proposed a data warehousing and mining approach for the discovery of unconventional petroleum ecosystems. A multidimensional star schema was used for the data warehouse. They attempt to make use of ontologies written for multiple dimensions to facilitate connections among unconventional petroleum ecosystems. The authors of [4] provide a reference architecture of sensor data warehousing. They make sure that the data is contextualized into the environment in which it was collected to avoid any misinterpretations. They further implemented the model on two cases studies. In [5], the authors propose an architecture to provide on-demand data warehousing to improve business intelligence. It is aimed at small and medium businesses that that a large volume of data but are unable to setup data warehousing infrastructure. The system is based on service-oriented architecture. In [6], the authors develop a data warehouse for an integrated blood donation management system. The historical blood donation data is stored in a centralized database for analytical processing. This would enable authorities to perform an informed decision on the location of blood donation drives. The authors further developed a Philanthropy Score assigned to the donors to quantify how good a citizen is. Another similar system is proposed in [7], here the authors develop a business intelligence system for education management. It enables schools to answer questions related to scoring, grading, demographics and visualize them using a dashboard. In [8], the authors propose a spatial based data warehouse using OLAP and SOLAP for road accident analysis. The design method is driven by the metadata that allows expressing the needs by taking into account the available data. The follow a supply-driven method, from source data extraction to spatial data warehouse model. Most data warehouse architectures discussed so far are not real-time. The data stored is updated on periodic intervals. The authors of [9] use a real-time data warehouse system. It allows decision makers to access and analyze very recent data to support real-time decision processes. However, complex queries take time to process, hence the authors propose a system to better manage materialized views and to deal with view selection as well as view maintenance problems in real-time data warehouse. They successfully developed an algorithm called, dynamic selection of materialized views (DynaSeV) which selects views from results of incoming queries under the execution time and the storage space constraints of the system. Another variant of real-time data warehouse is proposed in [10], here due to the problems related to the computation of large volumes of data, an ETL technique with zero latency is proposed, that works by constantly processing small chunks of data. Further techniques used to achieve zero latency are the use of logs, triggers, materialized views and timestamps.

3. Architecture

The architecture is designed depending on the business requirement. The subject matter considered here is Smartphone Sales. Star schema was chosen to model the data warehouse. The schema is depicted in figure 1.

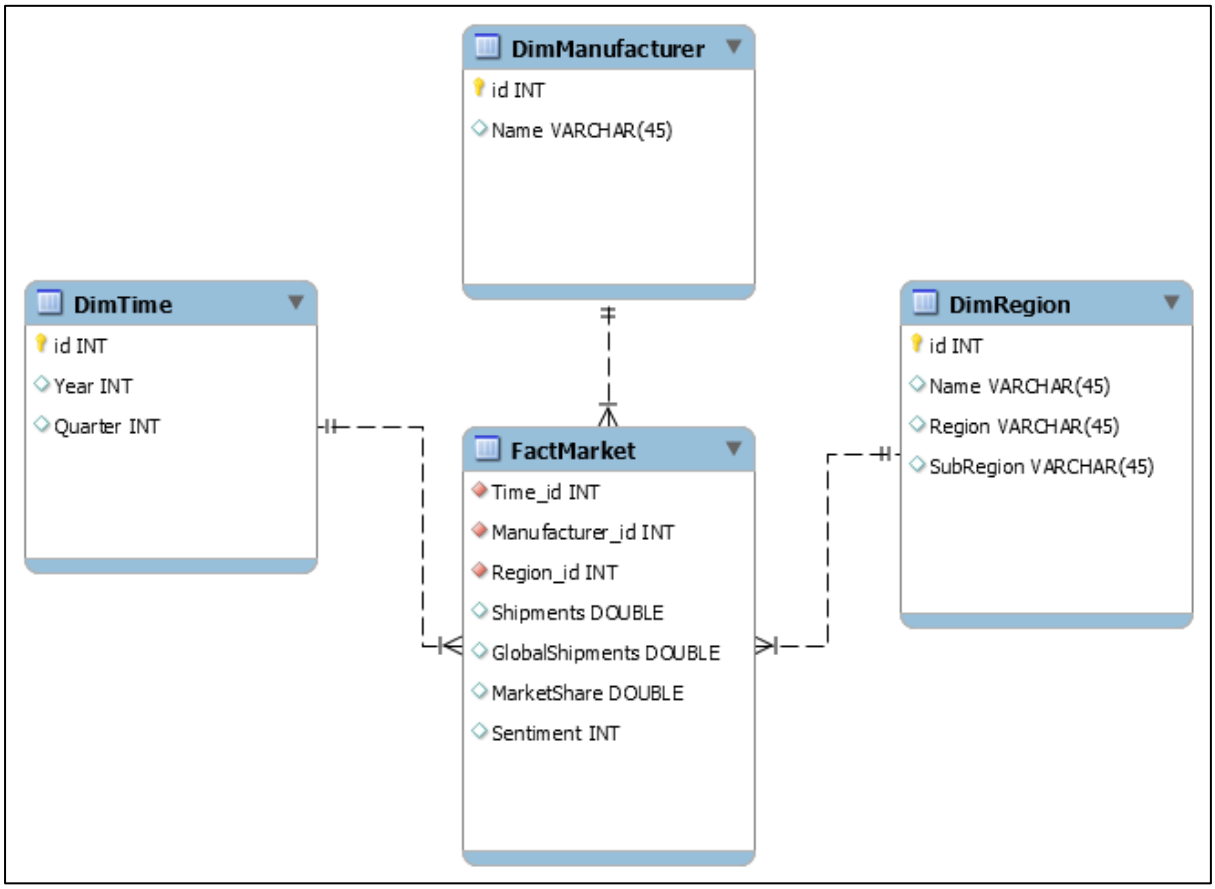

Figure 1: Data Warehouse Schema

A single fact table and three-dimension tables were identified. The fact table contains the following measurements: Shipments, Global Shipments, Market Share and Sentiment. The dimensions are chosen appropriately as per the subject matter. Manufacturer: This dimension contains the manufacturer name. Region: This dimension represents the sale location, three levels of hierarchy are provided (country, subregion, region). Time: This dimension represents the chronological nature of the data, two levels of hierarchy are provided (quarter, year). With the prepared schema, we move ahead to design the architecture of the warehouse. The top-down approach is used to design the data warehouse. The first step is organizing the data sources. Multiple heterogenous data sources are used. The data maybe present in existing operational systems, flat files, and other external sources. For this paper data was fetched from the following sources: Statcounter.com: Provides vendor-wise and region-wise smartphone market share. Statista.com: Provides vendor-wise smartphone shipment figures. Youtube.com: Provides user comment data from sentiment analysis. Restcountries.eu was used as an additional source to provide hierarchy data for the region dimension. The architecture used for the smartphone market is depicted in figure 2. It starts with fetching of data from the data

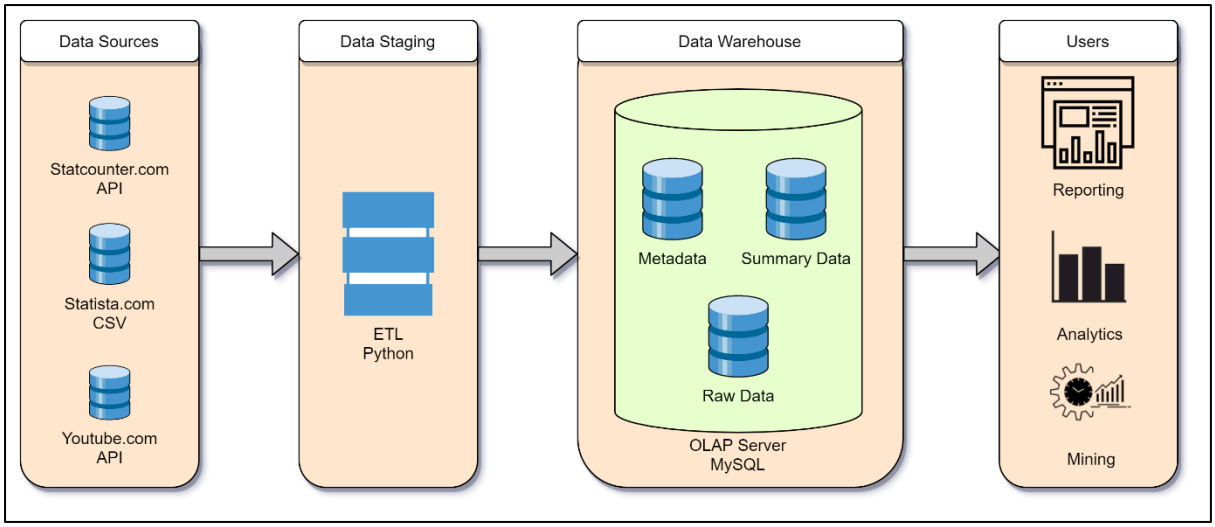

Figure 2: Data Warehouse Architecture

sources mentioned previously, followed by the data staging process which involves various transformations discussed in section 4. Next, the data will be loaded into the data warehouse and finally tools like Pentaho Business Intelligence Server are used to generate reports using MDX queries.

4. Extract-Transform-Load (ETL) process

The next step is Data Staging. The Extract-Transform-Load (ETL) process takes place to help integrate the data into the data warehouse from the various sources. It involves extracting the data from the data source, this might be a database, flat file, web service or other sources. For this paper the ETL process was mainly done in Python scripts. The data sources used were mainly CSV files and API endpoint resources. This requires usage of proper URL links to fetch the CSV files and API keys to authenticate to the respective API's. Data extraction from YouTube requires a few extra steps, as we need to determine the sentiment score from the raw text of a comment. Comments were searched for depending on the smartphone brand, country and time period in consideration, appropriate comments were fetched and VADER (Valence Aware Dictionary for Sentiment Reasoning) model was used for sentiment analysis. It is a lexicon and rule-based analysis tool that is best suited towards social media type of data, which makes it perfect for YouTube comments [11].

After extraction, the data is cleaned, this consists of detection and removal of invalid, duplicate or inconsistent data to help improve the quality and utility of the data before it is transferred to the data warehouse. In this step, tasks such as identification and removal of inconsistent data and approximating missing data are performed. The latter was performed on the data obtained from statista.com as it contained a few missing datapoints. In the next phase, the data is transformed. Here, the data is transformed into a format that can be stored and understood by the data warehouse, this requires standardizing a data type for the particular attribute being transformed. Depending on the number of sources, degree of heterogeneity and number of errors in the data, multiple data cleaning and transformation steps will be required. For this step, the tasks performed include, usage of regular

expressions to extract and format data; standardizing the date time formats, brand names, country names extracted from different sources. Also, handling of exceptional cases, such as no available YouTube comments are also performed in this step. In the final phase of ETL, the extracted and transformed data is loaded into the Data Warehouse; i.e., the fact and dimension tables. This was done using MySQL import. OLAP schema was also prepared using Pentaho Schema Workbench, and was published to the Pentaho Business Intelligence Server for analysis. At this point the Data warehouse is ready for use for analysis and report generation.

5. Experimental Results

The results were determined using the data from the previous steps. For analysis the following brands were considered: Apple, Samsung, Huawei, Xiaomi, Oppo, Lenovo. Along with data from the following countries: Australia, New Zealand, India, United Arab Emirates, Singapore, United Kingdom, Germany, United States, Brazil. Following four measures were available: shipments, global shipments, market share, sentiment. Results were retrieved from Pentaho Business Intelligence server using the JPivot view module. It provides an intuitive interface to visualize the data in the data warehouse. MDX queries are used to specify the data to be fetched.

Market share per region was per region was fetched using query in figure 3.

```
Select
NON EMPTY {[Measures].[Market Share]} ON COLUMNS,
NON EMPTY {[Region].[Americas], [Region].[Asia], [Region].[Europe],
[Region].[Oceania]} ON ROWS
from [Smartphone]
```

Figure 3: MDX Query for Market Share region-wise

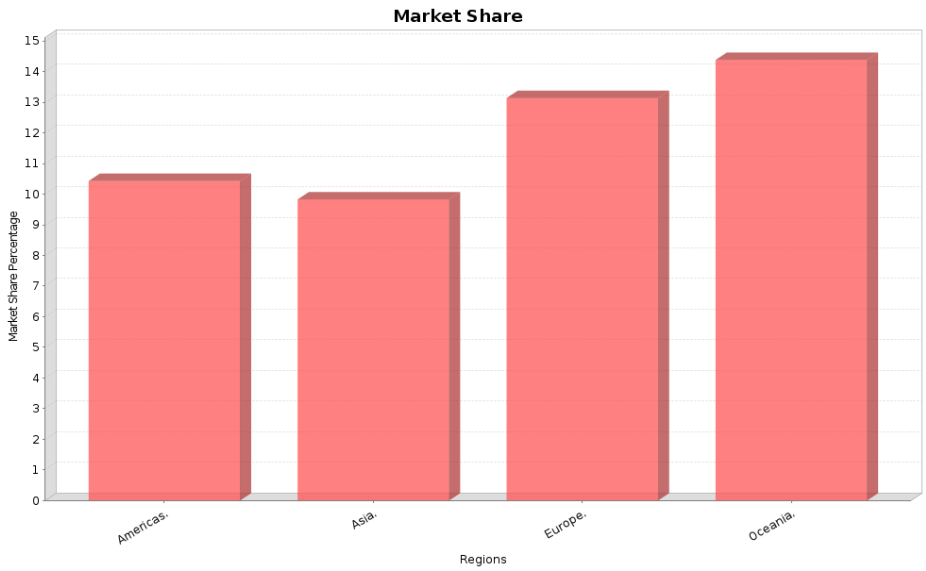

Figure 4: Market share region-wise

The query fetches the market share measure set as the column, and the regions set as the rows. Smartphone is the name of the cube. Vertical graph is used to represent the data. The graph in figure 5 shows more country specific data. This is achieved by drilling down on the

data from the previous query. Average is used as the aggregator for this purpose, as specified in the cube schema. Three levels of hierarchy are provided in the country dimension (e.g.: Americas, South America, Brazil).

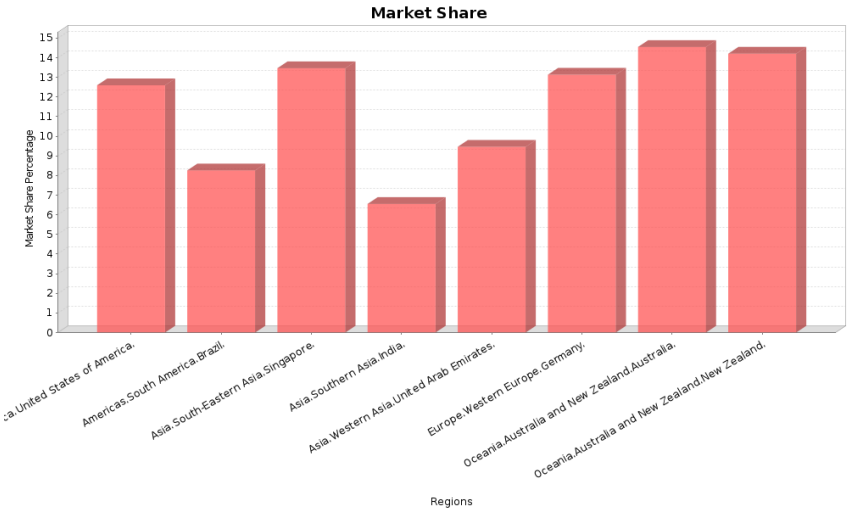

Figure 5: Market share country-wise

In the next query, shown in figure 6, the market share measure set is fetched as the column, however a crossjoin between the selected time set and the selected manufacturers set is fetched as the row. Crossjoin is a cross product between the two sets. This query enables us to get the market share of particular manufacturers over a selected period of time. Graph is shown in figure 7.

```
Select
NON EMPTY {[Measures].[Market Share]} ON COLUMNS,
NON EMPTY Crossjoin({[Time].[2020].[1]}, {[Manufacturer].[Apple],
[Manufacturer].[Huawei], [Manufacturer].[Lenovo], [Manufacturer].[Oppo],
[Manufacturer].[Samsung], [Manufacturer].[Xiaomi]}) ON ROWS
from [Smartphone]
```

Figure 6: MDX Query for Market Share manufacturer-wise during 2020 Q1

Figure 7: Market Share manufacturer-wise during 2020 Q1

In the next query shown in figure 8, The shipments for four selected brands in Q1 2020, is shown for two countries. This query makes use of a crossjoin in the column as well as then row. For the column, there is a crossjoin between the Shipments set and the selected manufacturers set. For the row, there is a crossjoin between the selected time period set (Q1 2020) and the selected countries set (USA, IN). The graph is shown in figure 9.

```
Select
NON EMPTY Crossjoin({[Measures].[Shipments]}, {[Manufacturer].[Huawei],
[Manufacturer].[Oppo], [Manufacturer].[Xiaomi], [Manufacturer].[Lenovo]}) ON
COLUMNS,
NON EMPTY Crossjoin({[Time].[2020].[1]}, {[Region].[Americas].[Northern
America].[United States of America], [Region].[Asia].[Southern Asia].[India]})
ON ROWS
from [Smartphone]
```

Figure 8: MDX query for Shipments from selected manufacturers in Q1 2020 for US and IN

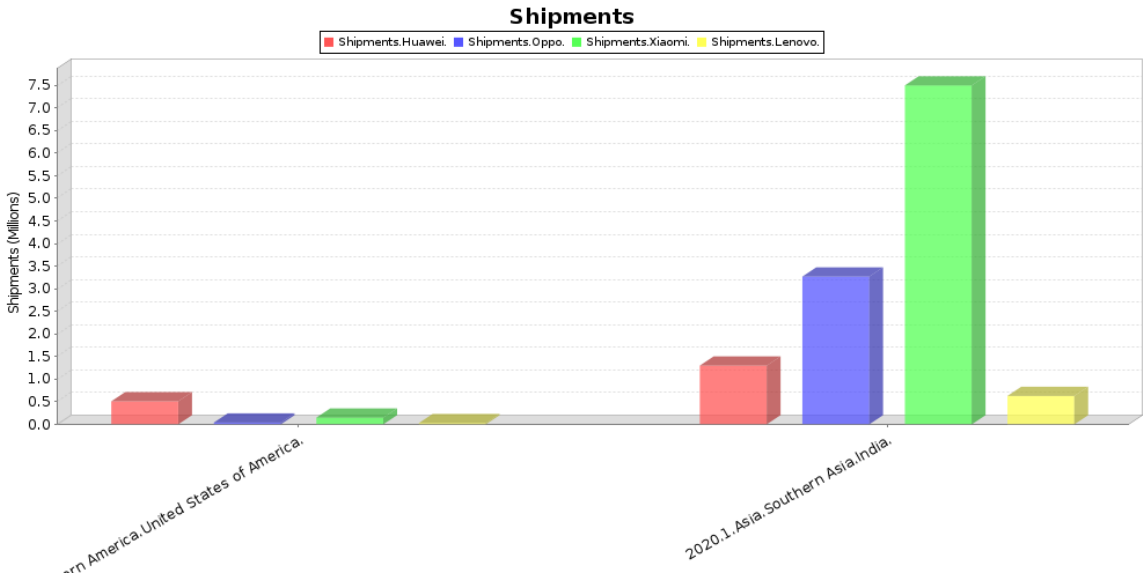

Figure 9: Shipments from selected manufacturers in Q1 2020 for US and IN

The next query shown in figure 10, is similar to the previous one, except that the column crossjoin consists of the countries set instead of the manufacturer set and the row crossjoin contains manufacturer set instead of countries set. Thus, depicting the flexibility of the MDX query language. Any of the dimensions can be selected for the row set or column set. The graph is shown in figure 11.

```
Select
NON EMPTY Crossjoin({[Measures].[Shipments]}, {[Region].[Americas].[Northern
America].[United States of America], [Region].[Asia].[Western Asia].[United
Arab Emirates]}) ON COLUMNS,
NON EMPTY Crossjoin({[Time].[2016], [Time].[2017], [Time].[2018],
[Time].[2019], [Time].[2020]}, {[Manufacturer].[Samsung],
[Manufacturer].[Apple]}) ON ROWS
from [Smartphone]
```

Figure 10: MDX query for shipments in US, AE for 2016-20 for Apple, Samsung

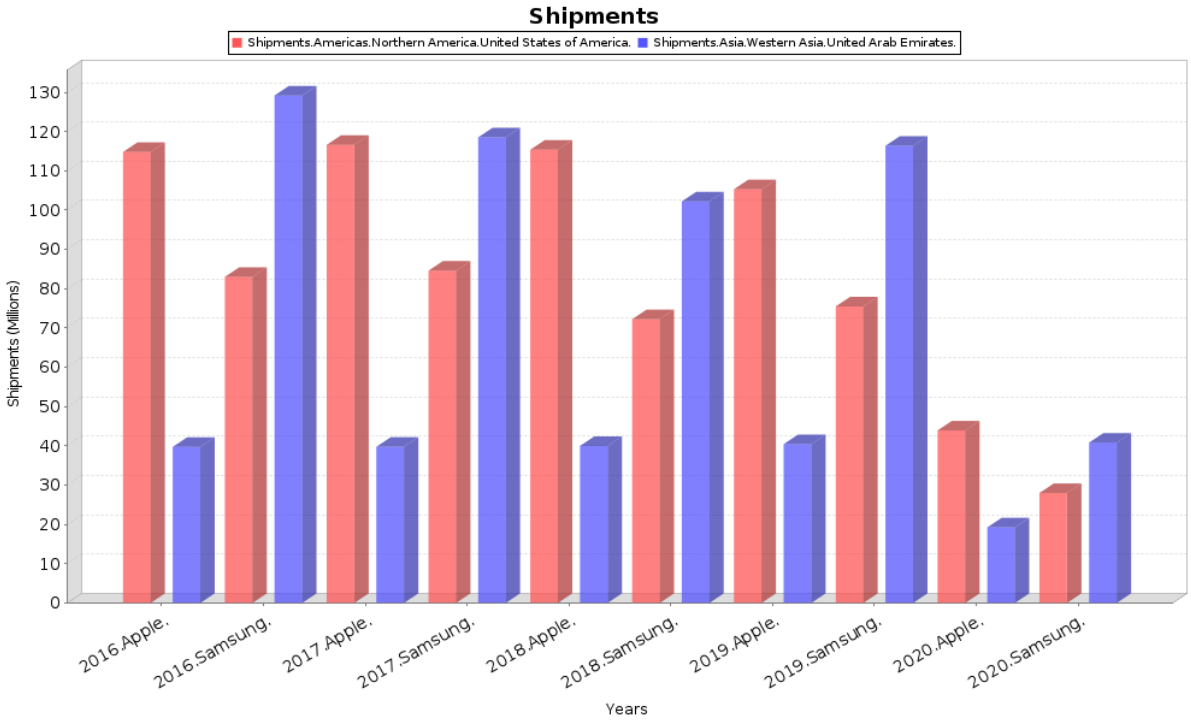

Figure 11: Shipments in US, AE for 2016-20 for Apple, Samsung

The next query shown in figure 12, makes use of two measures on the column, and three selected countries on the row. For the representation, pie chart by row was used, shown in figure 13.

```
Select
NON EMPTY {[Measures].[Shipments], [Measures].[Global Shipments]} ON COLUMNS,
NON EMPTY {[Region].[Asia].[South-Eastern Asia].[Singapore],
[Region].[Asia].[Southern Asia].[India], [Region].[Asia].[Western Asia].[United
Arab Emirates]} ON ROWS
from [Smartphone]
```

Figure 12: MDX query for shipments, global shipments from selected countries

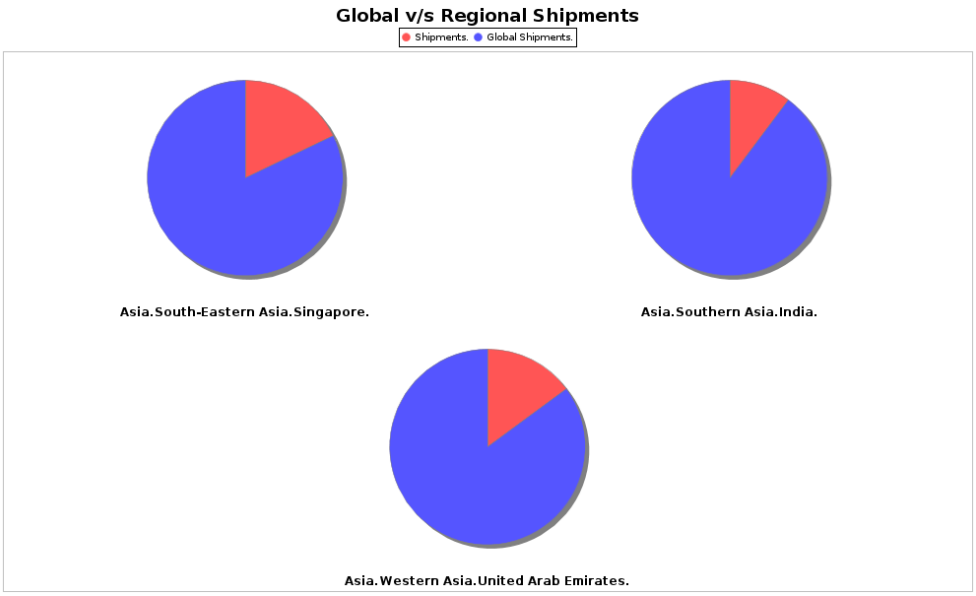

Figure 13: Pie chart comparing regional vs global shipments in selected countries

In the final query shown in figure 14, the sentiment measure is used. The column consists of a crossjoin between the sentiment measure set and three selected countries set. The row contains a crossjoin between the selected time set, and a manufacturer set. The line graph representation is shown in figure 15.

```
Select
NON EMPTY Crossjoin({[Measures].[Sentiment]}, {[Region].[Americas].[South
America].[Brazil], [Region].[Asia].[Western Asia].[United Arab Emirates],
[Region].[Europe].[Western Europe].[Germany]}) ON COLUMNS,
NON EMPTY Crossjoin({[Time].[2013], [Time].[2014], [Time].[2015],
[Time].[2016], [Time].[2017], [Time].[2018], [Time].[2019], [Time].[2020]},
{[Manufacturer].[Huawei]}) ON ROWS
from [Smartphone]
```

Figure 14: MDX query for sentiment in selected countries for selected time period from a selected manufacturer

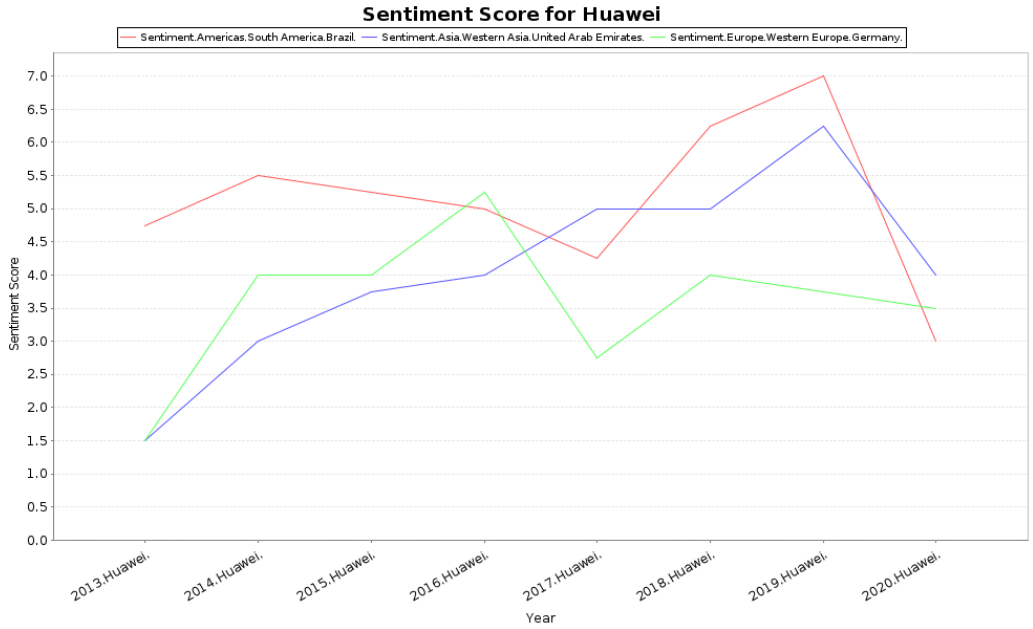

Figure 15: Line chart showing the sentiment for manufacturer over a period of time

6. Conclusion

Data warehouse allows us to get a historical view of the data. This proves very beneficial for analysis. In this paper, we have performed data warehousing and analysis for the Smartphone market. We began, with our schema depicting the various dimensions and fact tables, following the Star schema design followed by the architecture used to get the data from the data sources to the final reports stage. We then, described the detailed ETL process that was followed in this paper. Finally, in the Results section, we show the various types of MDX queries that were run and their corresponding output and graphical representations. In conclusion, Data warehouse is a very important aspect of Business Intelligence and is a quickly growing field, that is being used by a growing number of companies to help them get better insight into their operations.

References

[1] G. Maji and S. Sen, "A Data warehouse based analysis on CDR to depict market share of different mobile brands," 2015 Annual IEEE India Conference (INDICON), New Delhi, 2015, pp. 1-6, doi: 10.1109/INDICON.2015.7443706.

[2] G. Garani, A. Chernov, I. Savvas and M. Butakova, "A Data Warehouse Approach for Business Intelligence," 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Napoli, Italy, 2019, pp. 70-75, doi: 10.1109/WETICE.2019.00022.

[3] S. L. Nimmagadda, H. Dreher, P. A. Cardona Mora and A. Lobo, "Ontology based multidimensional data warehousing and mining of heterogeneous unconventional-reservoir ecosystems," 2013 11th IEEE International Conference on Industrial Informatics (INDIN), Bochum, 2013, pp. 535-540, doi: 10.1109/INDIN.2013.6622941.

[4] S. Dobson, M. Golfarelli, S. Graziani and S. Rizzi, "A Reference Architecture and Model for Sensor Data Warehousing," in IEEE Sensors Journal, vol. 18, no. 18, pp. 7659-7670, 15 Sept.15, 2018, doi: 10.1109/JSEN.2018.2861327.

[5] Y. Sharma, R. Nasri and K. Askand, "Building a data warehousing infrastructure based on service oriented architecture," 2012 International Conference on Cloud Computing Technologies, Applications and Management (ICCCTAM), Dubai, 2012, pp. 82-87, doi: 10.1109/ICCCTAM.2012.6488077.

[6] G. Maji, N. C. Debnath and S. Sen, "Data Warehouse Based Analysis with Integrated Blood Donation Management System," 2018 IEEE 16th International Conference on Industrial Informatics (INDIN), Porto, 2018, pp. 855-860, doi: 10.1109/INDIN.2018.8471988.

[7] A. S. Girsang, D. A. Sunarna, A. Syaikhoni and A. Ariyadi, "Business Intelligence for Education Management System," 2019 International Conference of Computer Science and Information Technology (ICoSNIKOM), Medan, Indonesia, 2019, pp. 1-6, doi: 10.1109/ICoSNIKOM48755.2019.9111559.

[8] N. Selmoune, K. Derbal and Z. Alimazighi, "Spatial Data Warehouse Multidimensional Design Approach and Geo-Decisional Tool for Road Accidents Analysis," 2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), Paris, France, 2019, pp. 1-8, doi: 10.1109/ICT-DM47966.2019.9032938.

[9] I. Hamdi, E. Bouazizi and J. Feki, "Dynamic management of materialized views in real-time data warehouses," 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Tunis, 2014, pp. 168-173, doi: 10.1109/SOCPAR.2014.7008000.

[10] C. R. Valêncio, M. H. Marioto, G. F. D. Zafalon, J. M. Machado and J. C. Momente, "Real Time Delta Extraction Based on Triggers to Support Data Warehousing," 2013 International Conference on Parallel and Distributed Computing, Applications and Technologies, Taipei, 2013, pp. 293-297, doi: 10.1109/PDCAT.2013.52.

[11] Gilbert, C. H. E., and Erric Hutto. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf. Vol. 81. 2014.